1N- 03　セグメントのクラスタリングと統合によるトピック分割

桂 薫　黄瀬 浩一　松本 啓之亮
大阪府立大学 工学部 情報工学科

1 はじめに

文書は複数の話題 (=マルチトピック) で構成されているものが多く，求める情報は文書の特定の話題 (=トピック) である場合がほとんどである．したがって文書をトピック毎に分割することは，情報へのアクセスを支援する一つの方法と考えられる [1].

従来から，トピック分割の手法としては様々なものが提案されている [2]．本稿では，これらのうち，Hearst らの手法 [3](以下，集約法と呼ぶ) にクラスタリング処理を導入することにより，類似性をより柔軟に判定する方法を提案する．

2 トピック分割

2.1 集約法とその問題点

集約法とは，文書を一定の単位で区切ったセグメントをもとに，隣り合うセグメントを統合することにより，トピック分割を実現する手法である．セグメントを統合するかどうかは，隣接するセグメント間の類似度により判定する．本稿では，以下のように単純化した集約法を比較対象として取り上げる．

セグメント $s_i(1 \leq i \leq S)$ に対して，索引語 $t_j(1 \leq j \leq L)$ を用いて，L 次元ベクトル $\boldsymbol{v}_i = (w_{i1}, \cdots, w_{ij}, \cdots, w_{iL})$ を考える．ここで，S はセグメント数，L は索引語数であり，

$$
\begin{aligned}
w_{ij} &= \sqrt{tf_{ij}} * \log(\frac{S}{df_j}) \\
tf_{ij} &= (\text{索引語 } t_j \text{のセグメント } s_i \text{内での出現回数}) \\
df_j &= (\text{索引語 } t_j \text{が出現するセグメント数})
\end{aligned}
$$

である．隣接するセグメント s_i, s_{i+1} 間の類似度は，

$$sim(s_i, s_{i+1}) = \frac{(\boldsymbol{v}_i, \boldsymbol{v}_{i+1})}{\|\boldsymbol{v}_i\| \|\boldsymbol{v}_{i+1}\|}$$

とする．そして，$sim(s_i, s_{i+1}) \leq T_1$ のときに，s_i と s_{i+1} は異なるトピックに属するものと判断する．

このような手法には，以下のような問題点がある．

- セグメントの長さを長くすると一つのセグメント中に複数のトピックが入り混じるため，分割精度が低下する．
- 一方，セグメントの長さを短くすると類似度を適切に計ることができない．すなわち，隣接するセグメントで共有する索引語が少なければ，同一トピックでも類似度が低くなる．

2.2 クラスタリングを用いたトピック分割

以上の問題点を解決するため，ここでは，セグメントのクラスタリングを導入する．手順を以下に示す．

1. セグメントを一つのクラスタとする．
2. 文書中で隣接するクラスタ間の類似度を求める．
   (a) 一方のクラスタに含まれるセグメントと，他方のクラスタに含まれるセグメントとの類似度を全て求める．
   (b) その内，最大の類似度をクラスタ間の類似度とする．
3. 類似度が最大の組合せのクラスタを併合し，一つのクラスタとする．
4. クラスタ数が一つだけになるまで 2,3 を繰り返すと，図 1のような木構造が得られる．
5. 以上の処理で得られた木構造を一定の類似度 T_2 以下の箇所で分割することにより，トピックを抽出する．

この手順は，隣接するクラスタを階層的クラスタリングにより順次結合していくものであり，ステップ 2により，隣接するセグメントが索引語を共有しなくても，隣接するクラスタに索引語を共有するセグメントが含まれていれば，類似度を計ることができる．

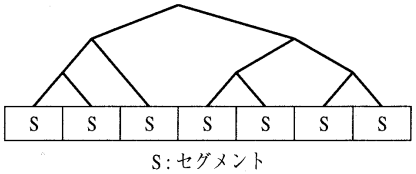

図 1: セグメントのクラスタリング

Topic Segmantation by Clustering and Merging of Segments
Kaoru Katsura, Koichi Kise
and Keinosuke Matsumoto
College of Engineering, Osaka Prefecture University

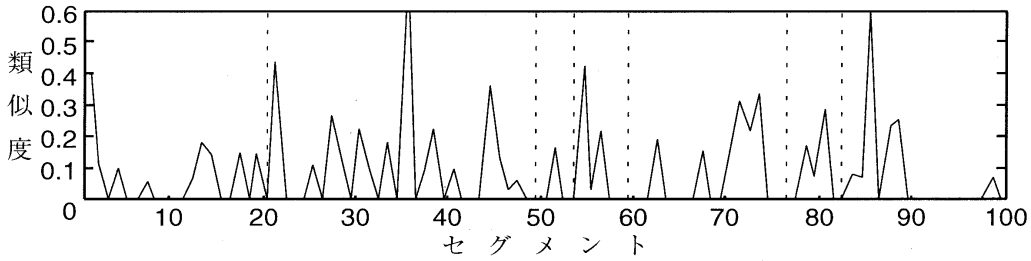

図 2: 集約法：10/5 付朝刊セグメント 1～100

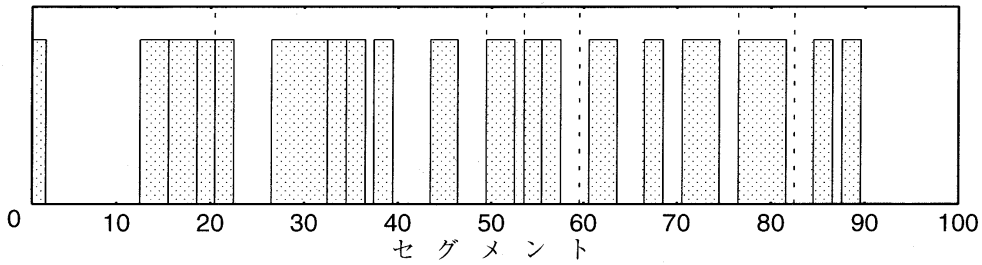

図 3: 本手法：10/5 付朝刊セグメント 1～100

3 実験

3.1 実験条件

実験サンプルとして，CD-毎日新聞 (94 年版) から 10 月 5 日付の朝刊の 197 記事を用いた．セグメントの長さは文単位とし，2659 個のセグメントを得た．索引語は記事を形態素解析 (juman3.5[4]) し，名詞と未定義語を採用した．ただしストップワードとして記号や「こと」「もの」等は除いた．閾値は $T_1 = 0, T_2 = 0.12$ とした．

3.2 実験結果と考察

集約法の処理例を図 2，本手法の処理例を図 3に示す．図 2, 図 3の横軸はセグメントの通し番号，図 2の縦軸は類似度を表す．記事の切れ目は，トピックの変わる箇所であり，図では縦線で表している．また，図 3では矩形がクラスタを表す．ただし，一つだけのセグメントからなるクラスタは表示していない．上図でセグメント番号 50 ならびに 77 辺りで，集約法が類似度 0 になっているにもかかわらず，本手法では正しくトピックの切れ目を検出できている．

正解をトピックの切れ目の位置として，以下の再現率，適合率を求めた．

$$\text{再現率} = \frac{\text{検出されたトピックの切れ目の内の正解数}}{\text{正しいトピックの切れ目の数}}$$

$$\text{適合率} = \frac{\text{検出されたトピックの切れ目の内の正解数}}{\text{検出されたトピックの切れ目の数}}$$

本手法では再現率 78.97%，適合率 22.06%の結果が得られた．集約法では再現率 76.41%，適合率 20.49%であった．

4 おわりに

本稿では，Hearst らの手法に階層的クラスタリングを導入することにより，トピック抽出を実現する手法を提案した．また，集約法と提案手法との比較実験を行い，その改善効果を確認した．

今後の課題には，LSI 等の導入による類似度の改良や，クラスタリング法の見直し等がある．

参考文献

[1] 長尾 眞: 自然言語処理, 岩波講座ソフトウェア科学 15, 岩波書店,pp.411-456(1996).

[2] 水野 浩之, 黄瀬 浩一, 松本 啓之亮: 窓関数を用いた部分テキスト検索—ベクトル空間法と出現密度法の比較—, 情処研資,NL135-2(2000).

[3] M.A.Hearst and C.Plaunt: Subtopic Structuring for Full-Length Document Access, Proc.of ACM Hypertext '93,pp.59-68(1993).

[4] ⟨URL:http://www-lab25.kuee.kyouto-u.ac.jp⟩.